Adaptive Dynamics and the Implementation Problem with Complete Information

Antonio Cabrales*
Universitat Pompeu Fabra and
University College London

July 19, 1996

Abstract

This paper studies the equilibrating process of several implementation mechanisms using naive adaptive dynamics. We show that the dynamics converge and are stable, for the canonical mechanism of implementation in Nash equilibrium. In this way we cast some doubt on the criticism of "complexity" commonly used against this mechanism. For mechanisms that use more refined equilibrium concepts, the dynamics converge but are not stable. Some papers in the literature on implementation with refined equilibrium concepts have claimed that the mechanisms they propose are "simple" and implement "everything" (in contrast with the canonical mechanism). The fact that some of these "simple" mechanisms have unstable equilibria suggests that these statements should be interpreted with some caution.

Key Words: Implementation; Bounded Rationality; Evolutionary dynamics; Mechanisms.

JEL classification: C72, D70, D78.

*This paper was written while the author was visiting University College London, and he would like to thank the members of the Economics department for their hospitality and encouragement. Thanks are also due to the helpful comments of Ken Binmore, Larry Samuelson, Richard Vaughan and a seminar audience at the London School of Economics, and to long conversations on implementation and economics with Luis Corchón, Roberto Serrano and Joel Sobel. The financial support of the European Union Human Capital and Mobility Program is gratefully acknowledged. Partial support was also obtained from Spain's Ministry of Education under grant PB93-0398

1 Introduction

The theory of implementation tries to address the problem of designing games (which in this literature are called *mechanisms*) whose equilibria satisfy certain socially desirable properties but which do not necessitate vast amounts of knowledge by the authorities to put them in place. Instead, these social arrangements should basically self police themselves, and the government should only make sure that the rules of the game are respected by the players.

In the last few years there have been impressive advances in the theory of implementation. As Sjöström (1994) points out, 'With enough ingenuity the planner can implement "anything"'. This "ingenuity" often involves the construction of complicated games and the use of refined equilibrium notions. As is often the case in economics, very little attention has been paid to the issue of how equilibrium is reached, and whether it is stable. The only exception we know of are the papers of Muench and Walker (1984) and de Trenqualye (1988), who study the local stability of the Groves and Ledyard (1977) mechanism. But we don't know of studies of more general mechanisms, either in Nash equilibrium or in some refinement of the Nash equilibrium concept; or studies of the issue of global convergence. This situation is worrisome given the importance of the issues at hand and the fact that the theory makes normative recommendations. It would not be sensible to apply these social engineering recipes without first thinking about whether real people will achieve the desired outcomes.

This paper tries to understand the effect of taking an adaptive (or evolutionary) dynamic approach for the implementation problem. The first result that we obtain is that the canonical mechanism for implementation in Nash equilibria (see Maskin 1977, Repullo 1987) has good dynamic properties under some additional assumptions about the outcomes of the social choice correspondence. If agents play the game repeatedly and replace the strategies they use with strategies that obtained higher payoffs in the past, the dynamics converge to the Nash equilibrium (so the social choice correspondence is implemented) and once the dynamics converge to the equilibrium, they stay there. According to Jackson (1992) 'A nagging criticism of the theory is that the mechanisms used in the general constructive proofs have "unnatural" features'. Moore (1992) also complains that the mechanisms for Nash implementation are 'highly complex -often employing some unconvincing device such as an integer game'. Our result shows that unsophisticated agents using very simple adjustment rules can reach the equilibrium of the mechanism, and therefore it is possible that the criticism is misplaced. On the other hand it may be that the critics are right. In this case we hope that our result encourages them to be more specific about the "complexity" they criticize.

The intuition for why convergence can be achieved with the canonical mechanism is simple. The structure of the general constructive mechanism is as follows. The agents have to announce a state of the world, an outcome and an integer. If all agents agree on a state and an outcome, the outcome is implemented. If one agent disagrees and proposes an

alternative, there is a test that the alternative has to pass. If it passes the test, the alternative outcome is implemented, otherwise it is not. A condition called *monotonicity* (Maskin 1977) ensures that an alternative will only be proposed if the agents are lying. The mechanism also specifies what happens when more than one agent disagrees. In these cases the mechanism gives the agent who proposed the largest integer her favorite outcome given the state of the world she announces. It is clear that no situation with more than one disagreer can be an equilibrium (if the best outcomes of the different agents are different), all agents will try to announce higher and higher integers to obtain their favorite outcomes. This use of integers is what the critics usually complain about.

Integer game constructions, however, help the process of convergence to equilibrium. Since all disagreement outcomes are bad for somebody and there is an incentive for the harmed to say the truth and obtain their favorite outcomes, they create a tendency towards equilibrium. As for the other states, monotonicity destabilizes agreements on something wrong but (when suitably strengthened) stabilizes agreements on the truth. Having understood that the "mission" of integer games is to make certain states unpleasant and to direct the dynamics to the right path, it is not difficult to see that the mechanism can be modified and the integer games can be substituted by something that plays their same role. To show this we construct an alternative mechanism that makes use of penalties and substitutes them for the integer game.

We also examine a mechanism that makes use of a more refined equilibrium concept. We show that although convergence to the equilibria of these games can be achieved, they are not very robust. The problem is that drift between strategies that have the same payoff as the equilibrium payoff can destabilize the equilibrium outcome. This result is far from being merely a theoretical curiosity. As Binmore and Samuelson (1996) point out, "the experimental evidence is now strong that one cannot rely on predictions that depend on deleting weakly dominated strategies", which is precisely what most of the mechanisms that use refined equilibrium concepts do. The mechanism we study, which is the one proposed by Abreu and Matsushima (1994), implements the social choice function in iteratively weakly undominated strategies. Besides being a good example of the literature on implementation with refined equilibrium concepts, it has an additional interest because it allows us to discuss the mechanism of Abreu and Matsushima (1992). This mechanism *virtually* implements the social choice function (that is, it implements the social choice function with arbitrarily high probability) in strategies that survive the iterative deletion of strictly dominated strategies. This would seem to be a good mechanism from a dynamic perspective, given that iteratively strictly dominated strategies are asymptotically eliminated for most adaptive dynamics (see Nachbar 1990, Samuelson and Zhang 1992 or Cabrales and Sobel 1992). The problem is that if the mechanism implements with very high probability the social choice function, then it will do so in iteratively strictly ϵ-*undominated* strategies, for ϵ very small. This implies that as the mechanism becomes more effective in doing its job, it becomes closer to the one in Abreu and Matsushima (1994) and thus it becomes open to the sort of instability problems which that mechanism has. We think that this trade-off between close implementability and stability needs to

be pointed out and we formalize it.

Section 2 describes the model and the dynamics we use. Section 3 studies the problem of Nash implementation with adaptive dynamics, both with the canonical mechanism and with an alternative that does not use integer games. Section 4 studies the dynamics of the mechanisms of Abreu and Matsushima (1994) and Abreu and Matsushima (1992).

2 The model and the dynamics

There is a set $I = 1, ..., n$ of agents, and the preferences of agent $i \in I$ are represented with a (Von Neumann-Morgenstern) utility function $v_i : A \times \Phi_i \to R$, where A is a finite set of alternatives and Φ_i specifies a finite set of possible utility functions. An element ϕ_i of Φ_i is meant to represent the preferences of agent i over A. A *preference profile* is a vector $\phi = (\phi_1, ..., \phi_n)$, where $\phi_i \in \Phi_i$. The set of possible preference profiles, denoted by S, is a subset of $\Phi = \times_{i \in N} \Phi_i$. Since we are concerned with environments with complete information, the preference profiles will be common knowledge among the agents.

A *social choice function* is a (possibly multi-valued) mapping $F : S \to A$, where $S \subset \Phi$ is the set of possible preference profiles. A *mechanism* is a pair (M, g), where $M = M_1 \times ... \times M_n$ and $g : M \to A$. M_i is the *message* space of agent i and g is the *outcome function*. A *mechanism* and a *preference profile* define a game.

Let $M_{-i} = M_1 \times ... \times M_{i-1} \times M_{i+1} ... \times M_n$. Given a *mechanism* (M, g) and a *preference profile* ϕ, we will say that m_i is a best response for player i, to $m_{-i} \in M_{-i}$ if $v_i(g(m_i, m_i), \phi_i) \geq v_i(g(m_i, m_i'), \phi_i)$ for all $m_i' \in M_i$. A message profile m is a *Nash equilibrium* (NE) if m_i is a best response to m_{-i} for all $i \in N$. Let $NE(\phi) = \{g(m) | m$ is a NE at $\phi\}$.

We say that a mechanism (M, g) *implements* a social choice function F in *Nash equilibrium* if for all $\phi \in S$, $F(\phi) = NE(\phi)$.

The main claim of this paper is that the implementation games can only be relevant for real individuals if one takes into account that an equilibrium of a game will be most of the time the result of some trial and error process of learning by the agents. In general there is no guarantee that an equilibrium will be the end product of such process. For this reason one should study if adjustment processes converge to some equilibrium of the implementation game. Furthermore, even if an equilibrium is attained it may be unstable, and stability would also be a desirable characteristic of a game form that is used to implement a social choice function.

We will assume now that the implementation game is played repeatedly by the agents and that they can use the information obtained in previous periods to modify their behavior in subsequent rounds of play. To keep the problem tractable we will make some assumptions about the way in which the play and the updating takes place.

Suppose that we have a population with L individuals playing the role of each agent i, so that the population has a total of Ln individuals. All individuals in the ith role are endowed with the same preferences ϕ_i. Let m_{ki} be the message sent by individual k in the role of agent i. A *population message profile* $s = (m_{11}, ..., m_{1n}, ..., m_{L1}, ..., m_{Ln})$ specifies a message for each individual in the population. A population message profile is *homogeneous* if m_{ki} is constant in k, that is, if all individuals in a certain role use the same message. Let the set S_{hom} be the set of all homogeneous message profiles. Let $s_i = (m_{1i}, ..., m_{Li})$ and $s_{-i} = (s_1, ..., s_{i-1}, s_{i+1}, ..., s_n)$. For a certain mechanism (M, g), preference profile ϕ, and message profile s let us denote $u_{ki}(s, \phi_i) = \sum_{k_1=1}^{L} ... \sum_{k_{i-1}=1}^{L} \sum_{k_{i+1}=1}^{L} ... \sum_{k_n=1}^{L} v_i(g(m_{k_1 1}, ..., m_{k_{i-1} i-1}, m_{ki}, m_{k_{i+1} i+1}, ..., m_{k_n n}, \phi_i) \frac{1}{L^{n-1}}$.

This utility function would obtain, for example, with uniform random matching, or (subject to renormalization) if every individual played with everybody else in the population, before they could change their behavior.

We will say that message $m_{ki} \in M_i$ improves upon strategy m'_{ki} given the message profile s if $u_{ki}(m_{ki}, s_{-i}, \phi_i) \geq u_{ki}(m'_{ki}, s_{-i}, \phi_i)$. We say that m_{ki} is a *best response* to s_{-i} if m_{ki} improves upon m'_{ki} for all $m'_{ki} \in M_i$. Let $\epsilon > 0$. We say that m_{ki} is an *ϵ-improvement* upon m'_{ki} if $u_{ki}(m_{ki}, s_{-i}, \phi_i) - u_{ki}(m'_{ki}, s_{-i}, \phi_i) > -\epsilon U$, where U is the maximum difference between payoffs for all $\phi \in \Phi$.

We assume that the population starts at some strategy profile and that before the game is played again one member of the population is allowed to change her strategy. The dynamics will be fully described when one identifies the probability with which each individual changes her message and the probability with which each message is chosen.

We will assume that the strategy shift will be made with a probability that depends exclusively on the present state of the population, and that with positive probability only one member of the population is allowed to change her strategy at any round of play. Besides this we will use a few other assumptions.

(D1) All individuals change their strategies with positive probabilities.

(D2) Any strategy that improves upon the strategy currently in use is adopted with positive probability.

(D3) A strategy that does not improve upon the strategy currently in use is adopted with zero probability.

Two alternatives for assumptions (D2) will be used in sections

(D4) Any strategy that is a best response to the present population profile is adopted with positive probability.

(D5) Any strategy that is an ϵ-improvement to the present population profile is adopted with positive probability.

It is important for our results that probabilities be bounded away from zero. This is implied by our assumptions when the strategy space and the population are finite. In section 3.1 the strategy spaces are infinite. We will use a modification of assumption (D2) in that section to deal with that problem. These properties make our dynamics very similar to the ones that Kim and Sobel (1995) use. The difference here is that (D1) is a strengthening of their (N1) assumption. When assumption (D4) is used the dynamics are closely related to the ones in Hurkens (1995), they are also a version for a discrete state space of the dynamics proposed by Gilboa and Matsui (1991). Assumption (D5) will be used in the discussion of virtual implementation.

Undoubtedly the dynamics described here are very crude, but one has to bear in mind that the games we analyze may have very large strategy spaces and the analysis could be very complicated without some drastic assumptions. We exploit a special characteristic that the mechanisms that are commonly used in the literature usually have. For many strategy profiles there are many agents who have lots of alternative strategies that yield the same payoff, and this payoff may be equivalent or even an improvement with respect to the one they are presently using. Some degree of variability in the response at those states will simplify convergence to an equilibrium. This variability is present if assumptions (D1) and (D2) (or the alternatives to D2) are satisfied. Assumption (D3) will make some equilibria absorbing states. These assumptions permit us to obtain clear-cut results in a relatively simple fashion. Other dynamics that are widely used in the literature may be less tractable. Nevertheless, a necessary further step would be to study the robustness of our conclusions to different assumptions about the dynamics.

3 Nash implementation

In this section we will argue that the mechanisms that have been used to implement social choice functions in Nash equilibria have good dynamical properties. In the first subsection we will show that the dynamics described in section (2) converge and are stable for the canonical mechanism. As we said in the introduction, this mechanism has been criticized for being "highly complex". One can argue that if agents that are as unsophisticated as the ones we model are able to converge to the equilibrium, the mechanism can hardly be called "highly complex". We also show that the features that have given the Nash mechanisms a bad name may not be necessary if the agents get to the equilibrium with our dynamics. That is the content of the second subsection.

3.1 The canonical mechanism (almost)

We say that F is *monotonic* if for all a, ϕ, ϕ', with $a \in F(\phi)$ and $a \notin F(\phi')$ there is an i and a' such that $v_i(a, \phi) \geq v_i(a', \phi)$ and $v_i(a', \phi') > v_i(a, \phi')$.

Let $b_i(\phi)$ be such that $v_i(b_i(\phi), \phi) \geq v_i(a, \phi)$ for all $a \in A$.

Monotonicity is a necessary and almost sufficient condition for Nash implementation. We will use somewhat stronger assumptions,

(N1) For all a, ϕ, ϕ', with $a \in F(\phi)$ and $a \notin F(\phi')$ there is an i and a' such that $v_i(a, \phi) > v_i(a', \phi)$ and $v_i(a', \phi') > v_i(a, \phi')$.

(N2) For all i, ϕ and $a \in F(\phi)$ there is $a' \in A$ such that $v_i(a, \phi) > v_i(a', \phi)$, and $v_j(a', \phi) \geq v_j(a, \phi)$ for $j \neq i$.

(N3) For all ϕ, $a \in F(\phi)$, for all j and for all $i \neq j$ $v_i(a, \phi) \geq v_i(b_j(\phi), \phi)$.

Let us denote the agent i in assumption (N1) $i(\phi, \phi')$[1], and a' be $a'(\phi, \phi')$. Agent $i(\phi, \phi')$ is often called the *test agent* and $a'(\phi, \phi')$ the *test outcome* in the implementation literature. Let us also denote by $a'_i(a, \phi)$ the outcome a' in assumption (N2).

Under our dynamics, all improving messages are chosen with positive probability. If the Nash equilibrium of the mechanism were such that some agent had more than one best response, it could be easily destabilized. To avoid this we will use two assumptions, (N1), which demands that the test outcome be a strict improvement over the "status quo" and (N2) by which it is always possible to punish a dissenter who has no reason to dissent (she is not a test agent) and also make everybody else better off. This will be possible if, for example, there is a private good and one can levy fines from one agent in terms of the private good and distribute the proceeds among everybody else. (N3) says that the social choice outcomes can be no worse for all agents $i \neq j$ than the favorite outcome for agent j. Suppose, for example, that there is a private good and consuming zero units of this good is very bad for any agent. Assume also that the favorite outcome for agent j assigns her all the group's endowment of the private good, and that all outcomes of the social choice function assign a positive amount of the private good to all agents. In these conditions assumption (N3) would be satisfied. Since the mechanism involves that for certain message profiles an agent gets her favorite outcome, this assumption prevents the dynamics from getting stuck at such a message profile.

We will use a slight variation of the canonical mechanism for implementation in Nash equilibria, as described, for example in Repullo (1987).

[1] If there is more than one agent i which satisfies the condition, let $i(\phi, \phi')$ be the one with the lowest index.

Let $M_i = A \times S \times N$, so that each individual announces an outcome, a preference profile, and a positive integer; and $M = M_1 \times ... \times M_n$, and let members of M_i and M be denoted m_i and m respectively. Let the first component of m_i, that is, the outcome announced by agent i be m_i^1 and the second component, the preference profile announced by agent i, be m_i^2. Let $i(m)$ be the individual who has the lowest index among those who announce the highest integer in the message profile m.

To define g, let's divide M into the following regions,

$$\begin{aligned}
D_1 &= \{m | \exists \phi \in S, a \in F(\phi) \text{ such that for all } i,\ m_i = (a, \phi, n_i),\ \text{for some } n_i \in N\} \\
D_2 &= \{m | \forall i \neq i(\phi, \phi'), m_i = (a, \phi, n_i) \text{and } m_{i(\phi,\phi')} = (a'(\phi, \phi'), \phi', n_{i(\phi,\phi')})\} \\
D_3 &= \{m | \forall i \neq j, m_i = (a, \phi, n_i) m_j \neq m_i \text{ and either } j \neq i(\phi, \phi') \\
&\quad \text{or } j = i(\phi, \phi') \text{ and } m_j \neq (a(\phi, \phi'), \phi', n_{i(\phi,\phi')})\} \\
D_4 &= \{m | m \notin D_1 \cup D_2 \cup D_3\}
\end{aligned}$$

$$g(m) = \begin{cases} a & \text{if } m \in D_1 \\ a'(\phi, \phi') & \text{if } m \in D_2 \\ a'_j(a, \phi) & \text{if } m \in D_3 \\ b_{i(m)}(m^2_{i(m)}) & \text{if } m \in D_4 \end{cases}$$

This mechanism can be described in the following way. If everybody agrees on an outcome and a state, then that outcome is implemented. If just one person disagrees with that announcement, and this person is the test agent and it announces the test outcome, then the test outcome is implemented. If just one person disagrees with that announcement, and this person is not the test agent (or it is the test agent but she does not announce the test outcome), then the dissenter is punished. If more than one person disagrees, then the outcome is the favorite one for the agent who announces the largest integer.

The only difference between this mechanism and the one in Repullo (1987) is that in this one we punish deviations from the equilibrium by agents other than the *test agent*, (and even punish announcements by the *test agent* which are not part of the test pair).

3.2 The dynamics of Nash implementation

The main result in this section is that with the dynamics defined in section 2 for the game that is defined by the mechanism in subsection 3.1 the population profile eventually induces one of the outcomes that the designer wants to implement with probability one, and that outcome is then implemented forever. In addition, if the initial state is homogeneous and none of the outcomes that the designer wants to implement are already being

implemented, all outcomes in the social choice function are implemented with positive probability.

As we said in section 2, we need to modify assumption (D2) to account for the fact that the strategy spaces are infinite since we need that some transition probabilities are bounded away from zero. The strategies that improve upon the one currently in use can be partitioned into a finite number of sets in which the strategies only differ with respect to the integer that is named. We assume (only in this section) that at least one strategy in each of these sets is used with a probability that is bounded away from zero.

Define the set $S_a = \{s | \exists \phi$, such that $\forall a \in F(\phi), i$, and $k, m_{ik} = (a, \phi, n_{ik})\}$. The set S_a is the set of population profiles where in all games played in the population the message profile is in the set D_1, and the outcome is a.

Proposition 1. Let the true preference profile be ϕ. Given dynamics that satisfy properties (D1), (D2), (D3), and given a social choice function that satisfies (N1), (N2), (N3);
(a) If $s(0)$ is a homogeneous population profile such that $s(0) \notin S_a$ for any $a \in \phi$, then for all $a \in \phi$ $P(\text{for some } t', s(t) \in S_a \forall t \geq t') > 0$.
(b) $P(\cup_{a \in F(\phi)} \{\text{for some } t', s(t) \in S_a, \forall t \geq t'\}) = 1$.

Proof: The proof will proceed through a series of lemmas. First we will show that for any population profile the population can become a homogeneous population profile with positive probability. Then we will show that a homogeneous population profile which does not implement any social choice function outcome can lead to all homogeneous profiles whose outcomes are social choice function outcomes, and finally we will show that a population profile in S_a, where everybody announces the true preference profile cannot exit that set.

Lemma 1: Let any $s(t')$. Then $P(s(t) \in S_{hom} \text{ for some } t > t') > 0$.

Proof: By the definition of $u_{ik}(s, \phi)$, $u_{ik_1}(s, \phi) = u_{ik_2}(s, \phi)$, if $m_{ik_1} = m_{ik_2}$, and therefore, for any given profile s_{-i}, the set of best responses for all individuals in role i is identical. Suppose that all individuals in role 1 are called upon to move in sequence. This is an event that has positive probability by property (D1). Since s_{-1} does not change along that sequence of moves, the set of best responses for agents in role 1 does not change. Suppose that they all change to the same best response m_1^*. This has positive probability by property (D2). There is also positive probability that an analogous succession of events that lead to all agents in role 1 to play m_1^* will lead to all individuals in role i to play some m_i^*, without an intervening chance for individuals in roles $j < i$ to change their strategies. This will lead then with positive probability to a homogeneous profile s with $s_i = (m_i^*, ..., m_i^*)$.□

Lemma 2: Let the true preference profile be ϕ and let $s(t') \notin S_a$ and $s(t') \in S_{hom}$. Then, for all $a \in F(\phi)$, $P(\text{for some } t > t', s(t) \in S_a) > 0$.

Proof: Given a homogeneous population message profile s, let $m(s) \in M$ be the message

profile associated to the homogeneous profile s. In the proof this lemma, all the strategy shifts that we will consider are shifts by all the individuals who are in the same role in sequence. Therefore expressions like "agent i will update her message", have to be understood as "all individuals in the role of agent i will update their message". The proof will proceed by dividing the possible initial states into a series of subsets.

Claim 1. For a given ϕ, if $s(t') \notin S_a$, for any $a \in F(\phi)$ and $s(t')$ in S_{hom} and $m(s(t')) \in D_1$, then $P(\text{for some } t > t',\ s(t) \in S_{hom} \text{ and } m(s(t)) \in D_4) > 0$.

Since $s(t') \in D_1$ and $s(t') \notin S_a$ all agents must be announcing a preference profile $\phi' \neq \phi$. By assumption (N1) and the definition of the mechanism, agent $i(\phi, \phi')$ can improve her payoff by announcing ϕ. Then with positive probability, by assumptions (D1) and (D2), agent $i(\phi, \phi')$ will have a chance to update and will choose to announce ϕ. After agent $i(\phi, \phi')$ changes her announcement , any agent $i \neq i(\phi, \phi')$ announcing state ϕ will move the message profile to a state in D_4. If at the same time she announces a high enough integer so that $i = i(m)$, then it will be advantageous to do so. Therefore this will happen with positive probability by (D2).□

Claim 2. Let $s(t')$ in S_{hom} and $m(s(0)) \in D_2$. Then $P(\text{for some } t > t',\ s(t) \in S_{hom} \text{ and } m(s(t)) \in D_4) > 0$.

If $m(s(t'))$ is in D_2, $m_i(s(t')) = (a', \phi', n_i)$ for all $i \neq i(\phi', \phi'')$. Any agent other than $i(\phi', \phi'')$ can move the message profile to D_4 by announcing a different outcome than a'. If she also chooses an integer high enough, and the true preference profile ϕ, she can obtain $b_i(\phi)$, which is a best response to $s(t')$. Assumptions (D1) and (D2) guarantee that this happens with positive probability. □

Claim 3. Let $s(t')$ in S_{hom} and $m(s(t')) \in D_3$. Then $P(\text{for some } t > t',\ s(t) \in S_{hom} \text{ and } m(s(t)) \in D_4) > 0$.

If $m(s(t'))$ is in D_3, $m_i(s(t')) = (a', \phi', n_i)$ for all $i \neq j$. Any agent other than j can move the message profile to D_4 by announcing a different outcome than a', and by choosing an integer high enough, and the true preference profile ϕ (which may or may not be equal to ϕ'), she can obtain $b_j(\phi')$, which is a best response to $s(t')$. Assumptions (D1) and (D2) guarantee that this happens with positive probability. □

Claim 4. Let the true preference profile be ϕ and $a \in F(\phi)$. Let $s(t)$ in S_{hom} and $m(s(t)) \in D_4$. Then $P(\text{for some } t' > t,\ s(t') \in S_{hom} \text{ and } m_i(s(t')) = (a, \phi, n_i)) > 0$.

If $m(s(t)) \in D_4$, there must be at least three different messages, and at least two of them have to be different from (a, ϕ). Call these reports, (a_1, ϕ_1), (a_2, ϕ_2), and the agents that send these messages i_1 and i_2 respectively. Let agents i with $i \neq i_1$ and $i \neq i_2$, replace their messages by (a, ϕ). These replacements still give rise to profiles in D_4 and they are best responses, provided that agents also announce the highest integer, thus assumptions (D1) and (D2) guarantee that they are sent with positive probability. Now we have to distinguish two cases.

(a) $i_1 = i(\phi, \phi_1)$, and $(m^1_{i_1}, m^2_{i_1}) = (a(\phi, \phi_1), \phi_1)$ and $i_2 = i(\phi, \phi_2)$, and $(m^1_{i_2}, m^2_{i_2}) = (a(\phi, \phi_2), \phi_2)$
(b) $i_1 \neq i(\phi, \phi_1)$, or $i_1 = i(\phi, \phi_1)$ and $(m^1_{i_1}, m^2_{i_1}) \neq (a(\phi, \phi_1), \phi_1)$; or $i_2 \neq i(\phi, \phi_2)$, or $i_2 = i(\phi, \phi_2)$ and $(m^1_{i_2}, m^2_{i_2}) \neq (a(\phi, \phi_2), \phi_2)$.
That is, either both i_1 and i_2 are test agents announcing test pairs, or one of them is not.

Let's start with case (b). Suppose i_1 fails to be a test agent announcing a test pair (this is without loss of generality). If agent i_2 replaced her message with (a, ϕ, n_{i_2}), this would lead to a profile in D_3, and the outcome would be $a'_{i_1}(a, \phi)$ (since there would be only one dissenter, and this dissenter would not be a test agent announcing a test pair). But by assumption (N2) agent i_2 prefers $a'_{i_1}(a, \phi)$ to a and, by (N3) this is an improvement over $b_{j(m)}$ for $j(m) \neq i_2$, which is the current outcome, and thus i_2 replaces her message with (a, ϕ, n_{i_2}) with positive probability. Once agent i_2 announces (a, ϕ, n_{i_2}) the outcome is $a'_{i_1}(a, \phi)$. If i_1 then announces (a, ϕ), the profile will be in S_a. By assumption (N2) this is a best response for i_1, since $a'_{i_1}(a, \phi)$ is worse than a for i_1, so i_1 will announce (a, ϕ) with positive probability.

In case (a) i_1 can replace her announcement by (a', ϕ), with $a' \neq a$. This preserves the state in D_4 and is a best response provided i_1 announces the highest integer. But now we are almost in the case (b) again. The only difference from case (b) is that i_1 is choosing the highest integer. But with positive probability some player other than i_1 and i_2 will move now and announce the highest integer, (which is a best response) and then the state will be like in case (b). At this point we can apply the argument in the previous paragraph to show that the transition to S_a has positive probability. □

Lemma 3: Let the true preference profile be ϕ and let $s(t) \in S_a$ for $a \in F(\phi)$ and $m^1_{ik}(t) = \phi$, for all i, k. Then $s(t') \in S_a$ for all $t' > t$.

Proof: If $s(t) \in S_a$, all message profiles for all possible matches are in D_1 and the outcome is a. The only replacements that can change something (since only one individual changes each time) will lead that game to a profile in D_2 or D_3. Since $m^1_{ik}(t) = \phi$, for all i, k assumptions (N1) and (N2) guarantee that these replacements do not mean an improvement for any agent, since a test agent announcing a test outcome for another profile ϕ' will obtain $v_i(a(\phi, \phi'), \phi) < v_i(a', \phi)$ by (N2) and any other deviating announcement (a', ϕ') obtains $v_i(a'_i(a, \phi), \phi) < v_i(a, \phi)$ by (N3). Since deviating messages produce strict losses, assumption (D3) guarantees that they will not be made. □

The combination of Lemmas 2 and 3 establishes part (a) of Proposition 1, since Lemma 2 shows that from any homogeneous profile the population reaches any S_a (for $a \in F(\phi)$) with positive probability and Lemma 3 shows that once the population is in S_a it never leaves the set. With the addition of Lemma 1 we have that from any state there is a probability $\epsilon > 0$ of reaching $\cup_{a \in F(\phi)} S_a$ and staying there forever in a number of steps smaller than some fixed and finite k. So the probability of not reaching $\cup_{a \in F(\phi)} S_a$ in kn steps is ϵ^{kn}. Since $\lim_{n \to \infty} \epsilon^{kn} = 0$, part (b) follows.□

3.3 A little further away from the canonical mechanism

Proposition 1 shows that the canonical game is not a bad idea from a dynamic point of view. One may still question it, however, on the grounds that integer games and similar constructions are very strange and they may not be "realistically implementable" (is there enough time in the universe life span to describe any arbitrary integer?[2])

From a purely dynamic point of view the assumption that all strategies that improve weakly upon the presently used one are taken with positive probability seems suspicious, especially given that strategies that do not improve are used with probability zero.

The answer to these questions is that under some assumptions on the permissible preference profiles one can construct a mechanism that does not use integer games and satisfies the good dynamic properties of the canonical mechanism, even with more restrictive requirements on the dynamics.

To be more precise, assumption (D2) will be replaced by assumption (D4). As for the assumptions that the preferences have to satisfy, we will drop (N3) and will add two others.

(N4) There exists an outcome P such that for all i, a, ϕ, with $a \in F(\phi)$ $v_i(a, \phi) > v_i(P, \phi)$.

(N5) For all ϕ, ϕ', $i \neq i(\phi, \phi')$, $v_i(P, \phi) \geq v_i(a'(\phi, \phi'), \phi)$.

Assumption (N4) creates a punishment that is worse than anything the designer wants to implement for everybody. We think of this as a kind of perverse "status quo" to which the situation will revert if there is widespread disagreement among the agents. Assumption (N5) tells us that if the test agent of the monotonicity condition denounces the other members of the group, the test outcome is implemented, and this is at least as bad for the "liars" as the P outcome.

Let $M_i = A \times S$, each individual announces an outcome and a preference profile. As before, $M = M_1 \times ... \times M_n$, and members of M_i and M are denoted m_i and m respectively. The first component of m_i, that is, the outcome announced by agent i is m_i^1 and the second component, the preference profile announced by agent i, is m_i^2. To define g, we divide M into three regions,

$$\begin{aligned} D_1 &= \{m | \exists \phi \in S, a \in F(\phi) \text{ such that for all } i,\ m_i = (a, \phi)\} \\ D_2 &= \{m | \forall i \neq i(\phi, \phi'), m_i = (a, \phi) \text{and } m_{i(\phi,\phi')} = (a'(\phi, \phi'), \phi')\} \\ D_3 &= \{m | m \notin D_1 \cup D_2\} \end{aligned}$$

[2] One should note that this criticism is not valid for mechanisms with 'modulo' games, which have the same dynamic properties as the canonical mechanism

$$g(m) = \begin{cases} a & \text{if } m \in D_1 \\ a'(\phi, \phi') & \text{if } m \in D_2 \\ P & \text{if } m \in D_3 \end{cases}$$

The explanation of this mechanism is similar to the canonical mechanism. If everybody agrees on an outcome and a state, that outcome is implemented. If just one person disagrees with that announcement, and this person is the test agent and she announces the test outcome, the test outcome is implemented. Otherwise, the outcome P is implemented.

One can see immediately that this mechanism does not implement the social choice function in Nash equilibria. Besides the equilibria that implement the social choice function, there are many other equilibria whose outcome is P. However, we will show that all equilibria, except the ones that implement outcomes of the social choice function, are unstable.

It turns out that with the dynamics defined in section 2 (even if we replace assumption (D2) for the harder to satisfy (D4)) the same conclusions obtained for Proposition 1 follow.

Define the set $S_a = \{s | \exists \phi, \text{ such that } \forall a \in F(\phi), i, \text{ and } k, m_{ik} = (a, \phi)\}$.

Proposition 2. Let the true preference profile be ϕ. Given dynamics that satisfy properties (D1), (D2), (D4), and given a social choice function that satisfies (N1), (N2), (N4), (N5);
(a) If $s(0)$ is a homogeneous population profile such that $s(0) \notin S_a$ for any $a \in \phi$, then for all $a \in \phi$ $P(\text{for some } t', s(t) \in S_a \forall t \geq t') > 0$.
(b) $P(\cup_{a \in F(\phi)} \{\text{for some } t', s(t) \in S_a, \forall t \geq t'\}) = 1$.

Proof: As with the proof of proposition 1, we will proceed through a series of lemmas.

Lemma 4: Let any $s(t')$. Then $P(s(t) \in S_{hom} \text{ for some } t > t') > 0$.

Proof: Same as Lemma 1. □.

Lemma 5: Let the true preference profile be ϕ and let $s(t') \notin S_a$ and $s(t') \in S_{hom}$. Then, for all $a \in F(\phi)$, $P(\text{for some } t > t', s(t) \in S_a) > 0$.

Proof:

Claim 5. If $s(t') \notin S_a$, for any $a \in F(\phi)$ and $s(t')$ in S_{hom} and $m(s(t')) \in D_1$, then $P(\text{for some } t > t', \ s(t) \in S_{hom} \text{ and } m(s(t)) \in D_3) > 0$.

Since $s(t') \notin S_a$ and $m(s(t')) \in D_1$ everybody must be announcing a preference profile $\phi' \neq \phi$. By assumption (N1), and by the definition of the mechanism, it is a best response for agent $i(\phi, \phi')$ to announce ϕ. Then with positive probability, by assumptions (D1) and (D4), agent $i(\phi, \phi')$ will have a chance to update and will choose to announce ϕ, which moves the profile to D_2. After agent $i(\phi, \phi')$ changes her announcement, any agent $i \neq i(\phi, \phi')$ announcing state ϕ will move the message profile to a state in D_3, and since

she can only move the game to D_3 or stay in D_2, and P is a better outcome for $i \neq i(\phi, \phi')$ by assumption (N5), announcing ϕ is a best response . Therefore this will happen with positive probability by (D4).□

Claim 6. Let $s(t')$ in S_{hom} and $m(s(t')) \in D_2$. Then P(for some $t > t'$, $s(t) \in S_{hom}$ and $m(s(t)) \in D_3) > 0$.

If $m(s(t'))$ is in D_2, $m_i(s(t')) = (a', \phi')$ for all $i \neq i(\phi', \phi'')$. Any agent other than $i(\phi', \phi'')$ can move the message to D_3 by announcing a different outcome than a', which is a best response to $s(t')$, by (N5). Assumptions (D1) and (D2) guarantee that this happens with positive probability. □

Claim 7. Let the true preference profile be ϕ and $a \in F(\phi)$. Let $s(t')$ in S_{hom} and $m(s(t')) \in D_3$. Then P(for some $t > t'$, $s(t) \in S_{hom}$ and $m_i(s(t)) = (a, \phi)) > 0$.

We can distinguish two cases.
(a) There exists some i_1 such that $m_{i_1} \neq (a, \phi)$ and either $i_i \neq i(\phi, \phi_1)$, for all $\phi_1 \in S$ or $m_{i_1} \neq (a(\phi, \phi_1), \phi_1)$.
(b) For all i, either $i = (a, \phi)$, or $i = i(\phi, \phi_1)$, for some $\phi_1 \in S$ and $m_{i_1} = (a(\phi, \phi_1), \phi_1)$.
That is, of all the agents that are not already announcing (a, ϕ), either there is one which is not a test agent announcing a test pair or all are test agents and announce test pairs.

In case (a) if all agents other than agent i_1 change their messages to (a, ϕ), the outcome is still P and it is a best response. If then i_1 changes her message to (a, ϕ), the outcome is a and this is a best response by assumption (N4).

In case (b) there must be at least two agents that are test agents announcing test outcomes, or $m(s(t))$ would not be in D_3. Let one of them be i_1, and let m_{i_1} change to (ϕ, a') for $a' \neq a$. This keeps the profile in D_3 and it is a best response. But now we are in case (a) □.

Lemma 6: Let the true preference profile be ϕ and let $s(t') \in S_a$ for $a \in F(\phi)$ and $m^1_{ik}(t') = \phi$, for all i, k. Then $s(t) \in S_a$ for all $t > t'$.

Proof: Like Lemma 3. □

The combination of Lemmas 5 and 6 establishes part (a) of Proposition 2, since Lemma 5 shows that from any homogeneous profile the population reaches any S_a (for $a \in F(\phi)$) with positive probability and Lemma 6 shows that once the population is in S_a it never leaves the set. With the addition of Lemma 4 we have that from any state there is a probability $\epsilon > 0$ of reaching $\cup_{a \in F(\phi)} S_a$ and staying there forever in a number of steps smaller than some fixed and finite k. So the probability of not reaching $\cup_{a \in F(\phi)} S_a$ in kn steps is ϵ^{kn}. Since $\lim_{n \to \infty} \epsilon^{kn} = 0$, part (b) follows.□

4 Refined and virtual Implementation

4.1 Implementation in iteratively undominated strategies

So far, we have only considered implementation in Nash equilibrium.What about more sophisticated equilibrium concepts? Since the seminal work of Moore and Repullo (1988), there has been considerable interest in implementation with more refined equilibrium concepts. The main advantage of these mechanisms is that the conditions for implementation are weaker. In particular monotonicity is no longer required. This is important since in economic environments implementing a single-valued social choice function and requiring monotonicity is equivalent to truthful implementation in undominated strategies (see Moore 1992).

By comparison, implementation in undominated strategies requires basically no restrictions. Abreu and Matsushima (1994) show that "any social choice function is exactly implementable in iteratively weakly undominated strategies", and Sjöström (1994) "in economic environments any social choice rule can be implemented in undominated Nash equilibria". An additional advantage of some of these mechanisms (notably those of Abreu and Matsushima (1994) and Sjöström (1994)) is that "integer games" or "modulo games" are not used.

The purpose of this section is to show that these advances should be viewed with some suspicion if we believe that equilibrium is the outcome of a learning process, since the adaptive dynamic process leads to undesired outcomes even asymptotically.

To focus the discussion we will concentrate on the mechanism proposed by Abreu and Matsushima (henceforth AM) (1994), but the results can be extended to other mechanisms based on refinements that have been proposed in the literature.

We will begin by introducing some notation and describing the mechanism.

The first thing to notice is that AM (1994) only consider single-valued social choice functions. Another important assumption is that there is a private good that can be used to levy (small) fines. Thus the utility function will be $v_i : A \times R \times \Phi_i \to R$. For simplicity we will use (as AM (1994) does) the quasi linear utility function $v_i(a, t, \phi_i) = u_i(a, \phi_i) + t_i$. Besides the outcome function $g(M)$ the mechanism specifies a *transfer rule,* $t = (t_i)_{i \in N} : M \to R^n$. The message space will be,

$$M_i = \Phi_i \times \Phi_{i+1} \times S \times \ldots \times S = M_i^{-1} \times M_i^0 \times M_i^1 \times \ldots M_i^K,$$

where K is an integer to be specified. By the lemma in AM (1992) we have that there exists a function $f_i : \Phi_i \to A$, such that for every $\phi_i \in \Phi_i$,

$$u_i(f_i(\phi_i), \phi_i) > u_i(f_i(\phi_i'), \phi_i) \text{ for all } \phi_i' \in \Phi_i / \{\phi_i\}.$$

Let $m = (m_1, \ldots, m_n)$, $m_i = (m_i^{-1}, m_i^0, \ldots, m_i^n)$, and $m^h = (m_1^h, \ldots, m_n^h)$. For any

message profile m, the outcome function is,

$$g(m) = \frac{e(m^0, \ldots, m^K)}{n} \sum_{i \in I} f_i(m_i^{-1}) + \frac{1 - e(m^0, \ldots, m^K)}{K} \sum_{h=1}^{K} \rho(m^h),$$

where for each $h = 1, \ldots, K$, we define $\rho : M^h \to A$ by

$$\rho(m^h) = \begin{cases} F(\phi) & \text{if } m_i^h = \phi \text{ for at least } (n-1) \text{ agents} \\ b & \text{otherwise, where } b \text{ is an arbitrary element of } A \end{cases}$$

and if we let ϵ be a small positive number to be specified later, and $\tilde{m}_0 = (m_n^0, m_1^0, \ldots, m_{n-1}^0)$, we define $e : M^0 \times \ldots \times M^K \to R$ by

$$e(m^h) = \begin{cases} \epsilon & \text{if } m_i^h \neq \tilde{m}_0 \text{ for some } h \in \{1, \ldots, K\} \text{ and some } i \in I \\ 0 & \text{otherwise} \end{cases}$$

The outcome function g is a lottery with the following characteristics. With a probability determined by the function e (which is nonzero when some agent's hth announcement differs from $\tilde{m}^0$) the favorite outcome of agent i, given her m_i^{-1}announcement, is selected with probability $1/n$. With probability $1 - e$ another lottery is chosen which gives equal weight to the K outcome functions given by the functions $\rho(m^h)$. This function says that if all but one of the m_i^h announcements coincide on ϕ, then $F(\phi)$ is implemented, otherwise an arbitrary outcome b is implemented.

To finish the determination of the mechanism the penalty function has to be specified. Let γ, ξ, η be small positive numbers to be specified later. Three possible penalties are specified for each player i.

1. γ if his zeroth announcement differs from player $(i+1)$'s minusoneth announcement.
2. ξ if his hth announcement $(h \geq 1)$ is the *first* to differ from $\tilde{m}^0$. All players who are first to deviate are punished.
3. η if his hth announcement is the only one to differ from the other players' hth announcements.

We will now give names to the fines

$$\tau(m_{i+1}^{-1}, m_i^0) = \begin{cases} -\gamma & \text{if } m_{i+1}^{-1} \neq m_i^0 \\ 0 & \text{otherwise} \end{cases}$$

$$d_i(m^0, \ldots, m^K) = \begin{cases} -\xi & \text{if } m_i^h \neq \tilde{m}_0 \text{ and } m_j^{h'} = \tilde{m}_0 \text{ for some } h \in \{1, \ldots, K\} \text{ and} \\ & \text{some } i \in I, \text{ all } j \in I, \text{ and all } h' \in \{1, \ldots, h-1\} \\ 0 & \text{otherwise} \end{cases}$$

$$\mu_i^h(m^h) = \begin{cases} -\eta & \text{if for some } \phi,\ m_i^h \neq \phi, \text{ but } m_j^h = \phi \text{ for all } j \in I/\{i\}, \\ 0 & \text{otherwise} \end{cases}$$

The total fine is thus $t_i(m) = \tau(m_{i+1}^{-1}, m_i^0) + d_i(m^0, \ldots, m^K) + \sum_{h=1}^K \mu_i^h(m^h)$.

To finish with the description of the implementation game we need to define the constants K, ϵ, η, ξ and γ. To do this define first,

$$E_i(\phi_i) = \max_{m^{-1} \in M^{-1}, m^h \in M^h} \left| \frac{1}{n} \sum_{j \in I} u_i(f_j(m_j^{-1}), \phi_i) - u_i(\rho(m^h), \phi_i) \right|$$

$$D_i(\phi_i) = \max_{m^h \in M^h, \bar{m}_i^h \in M_i^h} \left\{ u_i(\rho(m^h), \phi_i) - u_i(\rho(m_{-i}^h, \bar{m}_i^h), \phi_i) \right\}$$

Fix ϵ (small) and K (large) and choose η, ξ and γ to satisfy

Assumption AM1

$$\eta > \epsilon E_i(\phi_i)$$

$$\xi > \frac{1}{K} D_i(\phi_i) + \eta$$

$$\gamma > \epsilon E_i(\phi_i) + \xi$$

With these three inequalities AM (1994) show the following lemmas,

Lemma 7. Under assumption AM1. Let any m_i, and $\bar{m}_i = (\phi_i, m_i^0, \ldots, m_i^K)$, then for all m_{-i},

$$v_i(g(\bar{m}_i, m_{-i}), t(\bar{m}_i, m_{-i}), \phi_i) \geq v_i(g(m), t(m), \phi_i)$$

Lemma 8. Under assumption AM1. For all m with $m_i^{-1} = \phi_i$ for all i if we let $\bar{m}_i = (\phi_i, \phi_{i+1}, m_i^1 \ldots, m_i^K)$

$$v_i(g(\bar{m}_i, m_{-i}), t(\bar{m}_i, m_{-i}), \phi_i) > v_i(g(m), t(m), \phi_i)$$

Lemma 9. Under assumption AM1. For all m with $m_i^{-1} = \phi_i$, $m_i^0 = \phi_{i+1}$ and $m_i^q = \phi$ for all $q \in \{1, \ldots, h-1\}$, if we let $\bar{m}_i^q = m_i^q$ for all $q \neq h$ and $\bar{m}_i^q = \phi$ then,

$$v_i(g(\bar{m}_i, m_{-i}), t(\bar{m}_i, m_{-i}), \phi_i) > v_i(g(m), t(m), \phi_i)$$

We now show that if the lemmas are true, dynamics like those described in section 2 will go with positive probability to a state where the social choice function is implemented.

Proposition 3. Let the true preference profile be ϕ. Given dynamics that satisfy properties (D1), (D2), (D3), if Lemmas 7, 8 and 9 are satisfied, $P(\text{for some } t, s(t) \in S_{F(\phi)}) > 0$.

Proof: For this proof we first need some lemmas.

Lemma 10: Let any $s(t')$. Then $P(\text{for some } t \geq t', s(t) \in S_{hom}) > 0$.

Proof: Same as Lemma 1. □.

Lemma 11: Let $s(t') \in S_{hom}$ Then $P(\text{for some } t \geq t', s(t) \in S_{F(\phi)}) > 0$.

Proof: Let $s(t') \in S_{hom}$. Then with positive probability the players will change their messages one by one so that $m_{ki}^{-1}(t^{-1}) = \phi_{ik}$ for all i, k and some $t^{-1} > t$. That is, the minusoneth announcement of all players will be their true preferences. This happens because by Lemma 7 announcing the agent's own type truthfully in the minusoneth position is weakly dominant so assumptions D1 and D2 guarantee this will happen with positive probability. Similarly Lemmas 8 and assumption D1 and D2 guarantee that with positive probability there is a $t^0 > t^{-1}$ such that $m_{ki}^{-1}(t^0) = \phi_i$, $m_{ki}^0(t^0) = \phi_{i+1}$ for all i, k and Lemma 9 and assumption D1 and D2 guarantee that there is a sequence of time periods, $t^h > t^q$ for $q < h$ such that $m_{ki}^{-1}(t^h) = \phi_i$ $m_{ki}^0(t^h) = \phi_{i+1}$ $m_{ki}^q(t^h) = \phi$ for all i, k and $q < h$ Let then $t' = t^K$ □

This shows that the mechanism of AM (1994) can lead to the social choice function to be implemented. Unfortunately, it is also possible to diverge from the equilibrium in which the social choice function is implemented.

Proposition 4. Let the true preference profile be ϕ. Given dynamics that satisfy properties (D1), (D2), (D3), if $s(t) \in S_{F(\phi)}$, then $P(\text{for some } t' \geq t, s(t') \in S_{F(\tilde{\phi})}) > 0$ for any $\tilde{\phi}$.

Proof: If $s(t) \in S_{F(\phi)}$, then if agent n changes m_n^{-1} to some $\phi_n' \neq \phi_n$, her payoff does not change by the definition of the mechanism. D1 and D2 guarantee that this happens with positive probability. Let $\tilde{\phi}$ be such that $\tilde{\phi}_n = \phi'$ and $\tilde{\phi}_i = \phi$ for all $i \neq n$. Through a series of claims we show that with positive probability the population message profile goes to $S_{F(\tilde{\phi})}$ that is, $F(\tilde{\phi})$ is implemented.

Claim 1. If $m^{-1} = \tilde{\phi}, m_i^0 = \phi_{i+1}$ for all $i \in I$ and $m_i^h = \phi$ for all $i \in I$ and $h \geq 1$, then

$$v_{n-1}(g(\bar{m}_{n-1}, m_{-(n-1)}), t(\bar{m}_{n-1}, m_{-(n-1)}), \phi_{n-1}) - v_{n-1}(g(m), t(m), \phi_{n-1}) < 0$$

where $\bar{m}_{n-1} = (\tilde{\phi}_{n-1}, \tilde{\phi}_n, \tilde{\phi}, \phi, \ldots, \phi)$

$$\begin{aligned} &v_{n-1}(g(\bar{m}_{n-1}, m_{-(n-1)}), t(\bar{m}_{n-1}, m_{-(n-1)}), \phi_{n-1}) - v_{n-1}(g(m), t(m), \phi_{n-1}) \\ &= -\gamma + x(\phi) - \left(-\eta + \tfrac{\epsilon}{n}\textstyle\sum_{i \in I} f(\tilde{\phi}_i) + (1-\epsilon)x(\phi)\right) < 0 \end{aligned}$$

where the equality follows from the definition of the mechanism and the inequality follows from Assumption AM1.□

Claim 2. If $m^{-1} = \tilde{\phi}, m_i^0 = \tilde{\phi}_{i+1}$ for all $i \in I$ and $m_i^1 \in \left\{\tilde{\phi}, \phi\right\}$ (with at least $m_{n-1}^1 = \tilde{\phi}$)

and $m_i^h = \phi$ for all $i \in I$ and $h \geq 2$, then

$$v_i(g(\bar{m}_i, m_{-i}), t(\bar{m}_i, m_{-i}), \phi_i) - v_i(g(m), t(m), \phi_i) < 0$$

where $\bar{m}_i = (\tilde{\phi}_i, \tilde{\phi}_i, \tilde{\phi}, \phi, \ldots, \phi)$

If $m_i^1 = \tilde{\phi}$ only for $i = n-1$,

$$\begin{gathered} v_i(g(\bar{m}_i, m_{-i}), t(\bar{m}_i, m_{-i}), \phi_i) - v_i(g(m), t(m), \phi_i) \\ = -\xi + \tfrac{\epsilon}{n}\textstyle\sum_{i\in I} f(\tilde{\phi}_i) + (1-\epsilon)x(\phi) - \left(-\eta + \tfrac{\epsilon}{n}\textstyle\sum_{i\in I} f(\tilde{\phi}_i) + (1-\epsilon)\left(\tfrac{K-1}{K}x(\phi) + b\right)\right) < 0 \end{gathered}$$

If $m_i^1 = \tilde{\phi}$ for more than 1 but less than $n-2$ individuals,

$$\begin{gathered} v_i(g(\bar{m}_i, m_{-i}), t(\bar{m}_i, m_{-i}), \phi_i) - v_i(g(m), t(m), \phi_i) \\ = -\xi + \tfrac{\epsilon}{n}\textstyle\sum_{i\in I} f(\tilde{\phi}_i) + (1-\epsilon)\left(\tfrac{K-1}{K}x(\phi) + b\right) \\ - \left(-\eta + \tfrac{\epsilon}{n}\textstyle\sum_{i\in I} f(\tilde{\phi}_i) + (1-\epsilon)\left(\tfrac{K-1}{K}x(\phi) + b\right)\right) < 0 \end{gathered}$$

If $m_i^1 = \tilde{\phi}$ for $n-2$ individuals,

$$\begin{gathered} v_i(g(\bar{m}_i, m_{-i}), t(\bar{m}_i, m_{-i}), \phi_i) - v_i(g(m), t(m), \phi_i) \\ = -\xi + \tfrac{\epsilon}{n}\textstyle\sum_{i\in I} f(\tilde{\phi}_i) + (1-\epsilon)\left(\tfrac{K-1}{K}x(\phi) + b\right) \\ - \left(-\eta + \tfrac{\epsilon}{n}\textstyle\sum_{i\in I} f(\tilde{\phi}_i) + (1-\epsilon)\left(\tfrac{K-1}{K}x(\phi) + x(\tilde{\phi})\right)\right) < 0 \end{gathered}$$

If $m_i^1 = \tilde{\phi}$ for $n-1$ individuals,

$$\begin{gathered} v_i(g(\bar{m}_i, m_{-i}), t(\bar{m}_i, m_{-i}), \phi_i) - v_i(g(m), t(m), \phi_i) \\ = -\xi + \tfrac{\epsilon}{n}\textstyle\sum_{i\in I} f(\tilde{\phi}_i) + (1-\epsilon)\left(\tfrac{K-1}{K}x(\phi) + x(\tilde{\phi})\right) \\ - \left(-\eta + \tfrac{\epsilon}{n}\textstyle\sum_{i\in I} f(\tilde{\phi}_i) + (1-\epsilon)\left(\tfrac{K-1}{K}x(\phi) + x(\tilde{\phi})\right)\right) < 0 \end{gathered}$$

where the equalities follow from the definition of the mechanism and the inequalities follow from Assumption AM1.□

Claim 3. If $m^{-1} = \tilde{\phi}, m_i^0 = \tilde{\phi}_{i+1}$ for all $i \in I$ and $m_i^h = \tilde{\phi}$ for all $i \in I$ and $h \leq p$, and $m_i^h = \phi$ for all $i \in I$ and $h > p$, then

$$v_i(g(\bar{m}_i, m_{-i}), t(\bar{m}_i, m_{-i}), \phi_i) - v_i(g(m), t(m), \phi_i) < 0$$

where $\bar{m}_i$ is such that $\bar{m}_i^{-1} = \tilde{\phi}_i, \bar{m}_i^0 = \tilde{\phi}_{i+1}$ and $\bar{m}_i^h = \tilde{\phi}$ for all $h \leq p+1$ and $\bar{m}_i^h = \phi$ for all $h > p+1$,

$$\begin{gathered} v_i(g(\bar{m}_i, m_{-i}), t(\bar{m}_i, m_{-i}), \phi_i) - v_i(g(m), t(m), \phi_i) \\ = -\gamma + \tfrac{\epsilon}{n}\textstyle\sum_{i\in I} f(\tilde{\phi}_i) + (1-\epsilon)\left(\tfrac{K-p}{K}x(\phi) + px(\tilde{\phi})\right) \\ - \left(-\eta + \tfrac{\epsilon}{n}\textstyle\sum_{i\in I} f(\tilde{\phi}_i) + (1-\epsilon)\left(\tfrac{K-p}{K}x(\phi) + px(\tilde{\phi})\right)\right) < 0 \end{gathered}$$

where the equality follows from the definition of the mechanism and the inequality follows from Assumption AM1.□

Claim 4. If $m^{-1} = \tilde{\phi}, m_i^0 = \tilde{\phi}_{i+1}$ for all $i \in I$ and $m_i^h = \tilde{\phi}$ for all $i \in I$ and $h \leq p$, $m_i^{p+1} \in \left\{\tilde{\phi}, \phi\right\}$ (with at least one i such that $m_i^{p+1} = \tilde{\phi}$) and $m_i^h = \phi$ for all $i \in I$ and $h > p+1$, then

$$v_i(g(\bar{m}_i, m_{-i}), t(\bar{m}_i, m_{-i}), \phi_i) - v_i(g(m), t(m), \phi_i) < 0$$

where $\bar{m}_i$ is such that $\bar{m}_i^{-1} = \tilde{\phi}_i, \bar{m}_i^0 = \tilde{\phi}_{i+1}$ and $\bar{m}_i^h = \tilde{\phi}$ for all $h \leq p+1$ and $\bar{m}_i^h = \phi$ for all $h > p+1$,

If $m_i^{p+1} = \tilde{\phi}$ only for 1 individual,

$$\begin{gathered} v_i(g(\bar{m}_i, m_{-i}), t(\bar{m}_i, m_{-i}), \phi_i) - v_i(g(m), t(m), \phi_i) \\ = -\xi + \tfrac{\epsilon}{n}\textstyle\sum_{i\in I} f(\tilde{\phi}_i) + (1-\epsilon)\left(\frac{K-p}{K}x(\phi) + \frac{p}{K}x(\tilde{\phi})\right) \\ -\left(-\eta + \tfrac{\epsilon}{n}\textstyle\sum_{i\in I} f(\tilde{\phi}_i) + (1-\epsilon)\left(\frac{K-p-1}{K}x(\phi) + \frac{p}{K}x(\tilde{\phi}) + \frac{1}{K}b\right)\right) < 0 \end{gathered}$$

If $m_i^{p+1} = \tilde{\phi}$ for more than 1 but less than $n-2$ individuals,

$$\begin{gathered} v_i(g(\bar{m}_i, m_{-i}), t(\bar{m}_i, m_{-i}), \phi_i) - v_i(g(m), t(m), \phi_i) \\ = -\xi + \tfrac{\epsilon}{n}\textstyle\sum_{i\in I} f(\tilde{\phi}_i) + (1-\epsilon)\left(\frac{K-p-1}{K}x(\phi) + \frac{p}{K}x(\tilde{\phi}) + \frac{1}{K}b\right) - \\ \left(-\eta + \tfrac{\epsilon}{n}\textstyle\sum_{i\in I} f(\tilde{\phi}_i) + (1-\epsilon)\left(\frac{K-p-1}{K}x(\phi) + \frac{p}{K}x(\tilde{\phi}) + \frac{1}{K}b\right)\right) < 0 \end{gathered}$$

If $m_i^1 = \tilde{\phi}$ for $n-2$ individuals,

$$\begin{gathered} v_i(g(\bar{m}_i, m_{-i}), t(\bar{m}_i, m_{-i}), \phi_i) - v_i(g(m), t(m), \phi_i) \\ = -\xi + \tfrac{\epsilon}{n}\textstyle\sum_{i\in I} f(\tilde{\phi}_i) + (1-\epsilon)\left(\frac{K-p-1}{K}x(\phi) + \frac{p}{K}x(\tilde{\phi}) + \frac{1}{K}b\right) \\ -\left(-\eta + \tfrac{\epsilon}{n}\textstyle\sum_{i\in I} f(\tilde{\phi}_i) + (1-\epsilon)\left(\frac{K-p-1}{K}x(\phi) + \frac{p+1}{K}x(\tilde{\phi})\right)\right) < 0 \end{gathered}$$

If $m_i^1 = \tilde{\phi}$ for $n-1$ individuals,

$$\begin{gathered} v_i(g(\bar{m}_i, m_{-i}), t(\bar{m}_i, m_{-i}), \phi_i) - v_i(g(m), t(m), \phi_i) \\ = -\xi + \tfrac{\epsilon}{n}\textstyle\sum_{i\in I} f(\tilde{\phi}_i) + (1-\epsilon)\left(\frac{K-p-1}{K}x(\phi) + \frac{p+1}{K}x(\tilde{\phi})\right) \\ -\left(-\eta + \tfrac{\epsilon}{n}\textstyle\sum_{i\in I} f(\tilde{\phi}_i) + (1-\epsilon)\left(\frac{K-p-1}{K}x(\phi) + \frac{p+1}{K}x(\tilde{\phi})\right)\right) < 0 \end{gathered}$$

where the equalities follow from the definition of the mechanism and the inequalities follow from Assumption AM1.□

The claims show that $S_{F(\tilde{\phi})}$ is attained with positive probability because they show a series of changes in the messages, all of which are improving. Thus assumptions D1 and D2 guarantee that the sequence will take place with positive probability.

We have shown that there is positive probability of a transition between $S_{F(\phi)}$ and $S_{F(\tilde{\phi})}$ where $\tilde{\phi}$ differs from ϕ only in $\phi'_n \neq \phi_n$. But if the $s(t) \in S_{F(\tilde{\phi})}$, it is costless for individual $n-1$ to change $m_{n-1}^{-1} = \phi'_{n-1} \neq \phi_{n-1}$. By applying analogs of Claims 1 through 4 we can then show that with positive probability there is a time t' such that $s(t') \in S_{F(\ddot{\phi})}$, where $\ddot{\phi} = (\phi_1, \ldots, \phi'_{n-1}, \phi'_n)$. If we iterate this argument, the result follows.□

4.2 Virtual implementation

The idea behind virtual implementation is that to obtain implementability results under weaker sufficient conditions one can relax the notion of implementation (instead of strengthening the equilibrium concept). After all, the planner may well be satisfied as long as the social choice function is implemented with a high probability. AM (1992) show that if the planner only requires that the social choice function is implemented with arbitrarily high probability, basically any social choice function can be implemented, even with such a simple solution concept as iterative strictly undominated strategies.

This result would appear to be very congenial with the spirit of this paper. Since the solution concept is iterative strictly undominated strategies, both convergence and stability would be expected not only under the dynamics of this paper, but in a variety of evolutionary and learning models (see Nachbar 1990, Samuelson and Zhang 1992 or Cabrales and Sobel 1992). There is a problem, however, if the planner wants to implement a social choice function which is ϵ-close to the original social choice function. In that case some of the dominated strategies which have to be eliminated for the process to converge are only ϵ-strictly dominated. In fact we will show that if the agents are basically indifferent between strategies that give them utilities that are ϵ-close, then the same instability problems of the mechanisms of the previous subsection are reproduced here.

Following AM (1992), we say that a social choice function x and y are ϵ-close if for all preference profiles x and y map to lotteries that are ϵ-close. A social choice function x is *virtually implementable* in iterative strictly undominated strategies if for all $\epsilon > 0$, there exists a social choice function y which is ϵ-close to x and which is exactly implementable in iterative strictly undominated strategies.

To make the presentation a little simpler, we will not use the same mechanism that AM (1992) use but a modification based on AM (1994). As before we use the quasi linear utility function $v_i(a, t, \phi_i) = u_i(a, \phi_i) + t_i$. Besides the outcome function $g(M)$ the mechanism specifies a *transfer rule,* $t = (t_i)_{i \in N} : M \to R^n$. The message space will again be,

$$M_i = \Phi_i \times \Phi_{i+1} \times S \times \ldots \times S = M_i^{-1} \times M_i^0 \times M_i^1 \times \ldots M_i^K,$$

Let $m = (m_1, \ldots, m_n)$, $m_i = (m_i^{-1}, m_i^0, \ldots, m_i^n)$, and $m^h = (m_1^h, \ldots, m_n^h)$. The only

change in the mechanism is that for any message profile m, the outcome function is now,

$$g(m) = \frac{\epsilon}{n}\sum_{i\in I} f_i(m_i^{-1}) + \frac{1-\epsilon}{K}\sum_{h=1}^{K}\rho(m^h),$$

where for each $h = 1,\ldots,K$, we define $\rho : M^h \to A$ as before and ϵ is a small positive number as in the definition of *virtual implementation.* The penalty functions are also as specified before.

Note that with the modification made in the mechanism Lemma 7 is now true with a strict inequality.

Lemma 12. Under assumption AM1. Let any m_i, and $\bar{m}_i = (\phi_i, m_i^0, \ldots, m_i^K)$, then for all m_{-i},

$$v_i(g(\bar{m}_i, m_{-i}), t(\bar{m}_i, m_{-i}), \phi_i) > v_i(g(m), t(m), \phi_i)$$

Proof: Trivial from the proof of our lemma 7, and the definition of the mechanism.□

Lemma 12, plus lemmas 8 and 9 implies that implementation is in iterative strictly undominated strategies. Note also that the function implemented now is not F exactly but it is ϵ-close to F. Since ϵ can be made arbitrarily small, this mechanism virtually implements F. Let's denote the social choice function that is actually implemented for each value of ϵ, F_ϵ.

Proposition 5. Let the true preference profile be ϕ. Given dynamics that satisfy properties (D1), (D2), and (D3), if Lemmas 7, 8 and 9 are satisfied, for all $s(0)$ there exists t' such that $P(\text{for all } t \geq t', s(t) \in S_{F_\epsilon(\phi)}) = 1$.

Proof: A straightforward modification of the proof of proposition 3 shows that with probability 1 there exists t' such that $s(t') \in S_{F_\epsilon(\phi)}$ and the message $m_i = (\phi_i, \phi_{i+1}, \phi, \ldots, \phi)$ is sent by all players. Lemmas 12, 8 and 9 show then that for all $\bar{m}_i \neq m_i$,

$$v_i(g(\bar{m}_i, m_{-i}), t(\bar{m}_i, m_{-i}), \phi_i) - v_i(g(m), t(m), \phi_i) < 0$$

so by Assumption D3 $P(\text{for all } t \geq t', s(t) \in S_{F_\epsilon(\phi)}) = 1$.□

So, as mentioned earlier, the mechanism proposed guarantees very easily convergence and stability to a message profile that implements the social choice function with arbitrarily high probability, under assumptions (D1), (D2) and (D3).

The problem arises if assumption (D2) is replaced by (D5). We can then show,

Proposition 6. Let the true preference profile be ϕ. Given dynamics that satisfy properties (D1), (D3), (D5) if $s(t) \in S_{F(\phi)}$, then $P(\text{for some } t' \geq t, s(t') \in S_{F(\tilde{\phi})}) > 0$ for any $\tilde{\phi}$.

Proof: If $s(t) \in S_{F(\phi)}$, then if agent n changes m_n^{-1} to some $\phi_n' \neq \phi_n$, her payoff does not change by more than ϵM by the definition of the mechanism. Thus, D1 and D5

guarantees that this happens with positive probability. The rest of the proof retraces the steps of proposition 4 closely.□

This result implies that the agents have to care about the outcomes of the implementation process orders of magnitude more than the planner to avoid the instability of the mechanism. While this may be justified under certain circumstances, it is by no means a foregone conclusion.

5 Conclusions

The main message of this paper is that thinking explicitly about the equilibrating process in the implementation problem can be a fruitful experience. For example, one possible interpretation of the criticism of complexity that is levied against some general mechanisms is that boundedly rational agents cannot successfully achieve the equilibrium of the game. If this were the right interpretation, the criticism would be wrong, which at the very least should challenge the critics into making the criticism more concrete.

A possible reply to this result could be that it doesn't matter very much since we know that there are mechanisms that implement more things without using unnatural mechanisms. Our answer to this is that these mechanisms are objectionable since under reasonable dynamics the desired outcomes are not stable.

We hope that both of these results encourage more work into the implementation problem using dynamic tools. An important question that should be answered is how sensitive are our conclusions to the dynamics postulated. We suspect that the negative result about implementation in undominated strategies is bound to be robust to modifications of the dynamics. The result about the Nash mechanism may be more sensitive. In particular, we have not answered the question about the speed of adjustment. Reaching the social equilibrium may be irrelevant if it takes a very long time. It is possibly here where the critics of "unnatural" mechanisms may find a defense for their positions.

References

D. Abreu and H. Matsushima (1992), "Virtual Implementation in Iteratively Undominated Strategies: Complete Information", *Econometrica*, 60, 993-1008.

D. Abreu and H. Matsushima (1994), "Exact Implementation", *Journal of Economic Theory*, 64, 1-19.

K. Binmore and L. Samuelson (1996), "Evolutionary Drift and Equilibrium Selection", Institute for Advanced Studies, Vienna, Working Paper 26.

A. Cabrales and J. Sobel (1992), "On the Limit Points of Discrete Selection Dynamics", 57, 407-420.

I. Gilboa and A. Matsui (1991), "Social Stability and Equilibrium", *Econometrica*, 59, 859-867.

T. Groves and J. Ledyard (1977), "Optimal Allocation of Public Goods: a Solution to the Free Rider Problem", *Econometrica*, 45, 783-809.

S. Hurkens (1995), "Learning by Forgetful Players", *Games and Economic Behavior*, 11, 304-329.

M. O. Jackson (1992), "Implementation in Undominated Strategies: A Look at Bounded Mechanisms", *Review of Economic Studies*, 59,757-775.

M. O. Jackson, T. R. Palfrey and S. Srivastava (1994), "Undominated Nash Implementation in Bounded Mechanisms", *Games and Economic Behavior*, 6, 474-501.

Y. G. Kim and J. Sobel (1995), "An Evolutionary Approach to Pre-Play Communication", *Econometrica*, 63, 1181-1193.

E. Maskin (1977), "Nash Implementation and Welfare Optimality", mimeo, Massachusetts Institute of Technology.

J. Moore (1992), "Implementation in Environments with Complete Information", in J. J. Laffont ed., *Advances in Economic Theory: Sixth World Congress*, Econometric Society.

J. Moore and R. Repullo (1988), "Subgame Perfect Implementation", *Econometrica*, 58, 1083-1099.

T. Muench and M. Walker (1984), "Are Groves-Ledyard Equilibria Attainable?", *Review of Economic Studies*, 50, 393-396.

J. Nachbar (1990), "Evolutionary Selection Dynamics in Games: Convergence and Limit Properties", *International Journal of Game Theory*, 19, 59-89.

R. Repullo (1987), "A Simple Proof of Maskin's Theorem on Nash Implementation", *Social Choice and Welfare*, 4, 39-41.

L. Samuelson and J. Zhang (1992), "Evolutionary Stability in Asymmetric Games", *Jour-*

nal of Economic Theory, 57, 363-392.

T. Sjöström (1994), "Implementation in Undominated Nash Equilibria without Integer Games", *Games and Economic Behavior*, 6, 502-511.

P. de Trenqualye (1988), "Stability of the Groves and Ledyard Mechanism", *Journal of Economic Theory*, 46, 164-171.